A Spell Checker for a World Language: The New Microsoft's Spanish Spell Checker

Flora Ramírez Bustamante, Alfredo Arnaiz, Mar Ginés

Microsoft Speech and Natural Language Group
One Microsoft Way
Redmond, WA 98052
{florarb,alarnaiz,margines}@microsoft.com

Abstract

This paper reports work carried out to develop a speller for Spanish at Microsoft Corporation, discusses the technique for isolated-word error correction used by the speller, provides general descriptions of the error data collection and error typology, and surveys a variety of linguistic considerations relevant when dealing with a world language spread over several countries and exposed to different language influences. We show that even though it has been claimed that the state of the art for practical applications based on isolated word error correction does not offer always a sensible set of ranked candidates for the misspelling, the introduction of a finer-grained categorization of errors and the use of their relative frequency has had a positive impact in the speller application developed for Spanish (the corresponding evaluation data is presented).

1. Introduction

An important issue for any NLP system is how to deal with unknown words. Words can be unknown for different reasons: the word is not included in the lexicon used by the system because it is a proper name, a neologism, a foreign word, etc., or simply because it is a non-word, a misspelling. The purpose of spellers is to deal with such non-words.

The spell checker being presented here detects context-independent misspellings in general unrestricted texts, and provides isolated-word error correction, assisting the user by offering a set of candidate corrections that are close to the misspelled word. We define a misspelling as an unintended, orthographically incorrect representation of a word. The misspelling can differ from the intended word in the addition, omission, substitution and/or transposition of letters, space and punctuation marks. The spelling correction functionalities are based on string similarity and the well-known notion of *edit distance* (the number of insertions, deletions, substitutions and transpositions of two adjacent characters) as explained in Damerau (1964). The speller does not incorporate any special strategy for dealing with other types of unknown words as names (e.g. proper names, location names, etc.) or neologisms (see Toole, 2000; Vosse, 1992, respectively).

Once the speller recognizes a token in the user input as a non-word, it looks for a close match in the lexicon. Whether or not a match between a dictionary form and the user input is sufficiently good is determined via a numerical cost or score. After a score is computed for each possible suggestion, suggestions are ranked according to score. The suggestion with the lowest score is the best suggestion. Scores rate the standard four editing operations mentioned above plus word frequency (within a four-category scale).

2. Error probabilities

One part of the scoring process consists on assigning scores based on the probabilities of a given character or set of characters (i. e. error strings) being mistaken by another character or set of characters. These pairs of error or confusion strings are language dependent. They were extracted from the most common error/correction pairs from a typology (see below) and incorporated to the speller as additional editing operations. A probabilistic device was used to automatically predict the score of every editing operation. The device was trained on the confusion pairs and a large list of misspellings. This device predicted the best score of every confusion string via a genetic algorithm which induced random permutations on the scores until it ended up with the best score in accordance to an objective or fitness function (i.e. a type that quantifies the optimality of the score). These scores are used to improve the ranking of suggestions when the corresponding confusion strings are at stake.

The confusion pairs were extracted from the most probable error patterns found in a corpus (see below). Ninety-five particular confusion pairs were implemented. The resulting confusion pairs incorporated performance/typing errors due to stroke proximity, and cognitive errors (i. e. misconceptions about orthographic rules or lack of language knowledge). Because of these, the speller is able to provide sound suggestions for a variety of cognitive errors including errors involving irregular morphological paradigms of the type **andó-anduvo*,**volvido-vuelto,* and phonetic errors of the type **desición-decisión*.

3. The error typology

To ensure the coverage of relevant error types and also to provide data for the evaluation requirements, a typology of errors for Spanish was created based on the automatic and manual revision of corpora containing over 8 million words. These corpora include mainly written texts from Mexico and Spain, with a good spread of balanced lexical domains. The corpora comprise three different sets of texts. The first one, with about 4 million words, is a balanced set of edited and unedited texts. The second one is a highly edited corpus with about 2 million words. The third one, with about 2 million words, contains data from the Original Works Creation (OWC) sites collected for the Microsoft Office grammar checker for Spanish. These files were created to gather unedited text, and are not spell-checked or grammar-checked. The participants from

Madrid and Mexico City chose from several topics and wrote for about 30 minutes. Authors were only allowed to make minimal immediate corrections. The files have only been sentence-separated.

In order to find the spelling errors, the corpus was analyzed morphologically using a knowledge-based syntactic parser (similar to the parser which underlies the Office grammar checkers developed by Microsoft). The automatic morphological analysis produced a list of unknown words. Those words were manually revised. From the initial number of unknown words, we were able to derive around 76K error occurrences. A correction was assigned to each misspelled word based on the context, with the result that the same misspelled word could have more than one correction and therefore appear in more than one error pair. The typology contains around 27K unique error pairs. Every spelling error was classified as belonging to a more general class of errors (i. e. error type). The frequency of each error type and error pair occurrence was calculated.

We worked with an initial classification of 142 error types which gives an exhaustive account of the main four spelling error/editing operation types, encompassing not only characters but also other relevant features, such as accents, spaces, lower and upper case and the important classes of cognitive errors. This classification also considered multi-error misspellings, difference/distance in the position of error and correction, and number of instances of editing operations.

An analysis of the error typology revealed that

(a) almost 90% of errors are single error misspellings;

(b) omission is the most frequent error type, distantly followed by substitution, addition and transposition, in that order;

(c) it is necessary to look into not only characters but also diacritics and case:

(i) over 50% of the errors involve the omission of an accent;

(ii) substitution of lower case for upper case at the beginning of a proper noun word is one of the most common errors; and

(d) cognitive errors make up an important class of errors.

Table 1 shows the distribution of the most common spelling error patterns found in the corpus.

Error type	**%**
Omission of diacritics [Ex. **dia* vs. *día*]	51.5
Omission one character [Ex. **mostar* vs. *mostrar*]	6.8
Capitalization beginning word [Ex. **windows* vs. *Windows*]	6.2
Cognitive errors [Ex. **biene* vs. *viene*]	5.9
Addition of one character [Ex. **aereopuerto* vs. *aeropuerto*]	4.7
Substitution one character [Ex. **calavara* vs. *calavera*]	4.1
Addition of diacritics [Ex. **fuí* vs. *fui*]	2.9
Omission space [Ex. **esque* vs. *es que*]	2.0
Transposition one character no beginning word [Ex. **interpetración* vs. *interpretación*]	1.7
Capitalization whole word [Ex. **fifa* vs. *FIFA*]	1.3
Repetition of same letter by addition [Ex. **dirrección* vs. *dirección*]	1.1
Substitution of diacritic character [Ex. **informaciòn* vs. *información*]	0.5
Addition space [Ex. **bue na* vs. *buena*]	0.4
Other	0.24
Total single error misspellings	**89.34**
Multi-error: substitution (including diacritics) + addition, omission [Ex. **paguina* vs. *página*]	3.6
Multi-error: other [Ex. **Nesecitaria* vs. *Necesitaría*]	2.8
Multi-error: capitalization + addition, omission, substitution (including diacritics) [Ex. ** jose* vs. *Jose*]	1.1
Other	1.11
Total multi-error misspellings	**8.61**
Rest [*serguiovamos → seguro vamos*]	2.05

Table 1: Error type distribution in the corpus

In our typology we considered the special class of homophones in which the writer substitutes a phonetically correct but orthographically incorrect sequence of letters as cognitive errors (i.e. mistakes on similar phonetic pairs such as *b-v, s-x, c-s, ll-y* for instance, as in **archibo-archivo, *estención-extensión, *llendo-yendo*). This type of error is not only common in single error patterns but also in multi-error patterns. Additionally, the source of errors involving accents and case could be related as well to a misconception on the part of the writer. That would mean that more than 63% of the misspellings in our corpus would be of the cognitive nature.

The predominance of cognitive errors in Spanish can be explained by two factors: the Spanish orthographic system is closely based on phonetic patterns, and there is a tendency to resort to phonetic spellings when one does not know the spelling of a given word. In fact, the distribution of these errors varies depending on pronunciation differences between dialects: most Spanish dialects do not distinguish [s] and [θ], which explains the common confusion around “c”, “s”, and “z”. In Spanish, phonetic errors can be easily treated with the edit distance operations since the correction of most of them only implies one of such operations. However, the edit distance technique is bad suited to more complex phonographemic errors as Veronis (1988) claimed for French (e.g. **ippeautaineuz* for *hypoténuse*).

4. Lexical Coverage

The speller lexicon presently consists of a little over 1 million entries. Around 9% of these entries are special forms required by an enclitic runtime recognizer which allows the speller to validate and suggest verbal enclitic forms (e.g. *cántamela*, *cantáosla*, *preparándonoslo*). The use of this runtime enclitic form recognizer extends the coverage of the speller well over the actual number of entries.[1]

[1] The speller for Spanish based on the Unix tool *ispell* works with a lexicon of about 650,000 inflectional entries including enclitic forms, as it is reported in Rodríguez *et al.* (1995, 1996).

The lexicon also covers geographical names, names of persons and organizations, abbreviations, acronyms, and diminutive forms (mainly, adjectives and nouns) based on corpus frequency.[2] The addition of these forms was based on a process of harvesting of the most frequent unknown words found in a corpus of 900 million words. For instance, the diminutive forms were generated taking into consideration the 50K most frequent words. This produced a list of approximately 90K full/inflected diminutive forms.

Additionally, there are two other classes of words—specific to the functioning of the speller—worth mentioning, these are masking and restricted words. First, strictly speaking, *masking words* are legal words that may hide a spelling error. They are pairs of words in which both are potentially a correction of each other (e.g. *avía-había*, *bes-vez*, *como-cómo*, *coro-corro*). For most of these pairs, the consideration of each word frequency in conjunction with the frequency of error for one or both of them usually guides the decision of converting one of them into a non-word to be able to capture an error over the other. Thus, from the examples above, only *avía* and *bes* were labeled *masking*. For the other two cases, the frequency of the words overrode the frequency of the error; hence, both pairs remained untouched in the lexicon.[3] Second, *restricted words* are forms that for one reason or another are removed from the suggestion list. That is, the speller recognizes them, but do not use them as suggestions for corrections. This type of words were used mainly for two purposes: (i) eliminate offensive/sensitive words from the suggestion list to avoid a negative reaction from some users, and (ii) recognize forms that are widely spread and commonly used but are not normatively sanctioned (or typographically available): e.g. *Ma.* for *M.^{a}*.

During the analysis of the corpus, relevant coverage questions arose. Most of these issues relate to the particular case of Spanish, but are extensible to any other major language in a similar situation:

1. Which unfound words are candidates for inclusion in our lexicon, and then what should be the lexical coverage given that the spell checker is addressed to general Spanish speakers spread all over the world?
2. What should be done with restricted words in a language with so many dialects?
3. Which is the role of prescriptive lexical sources within a corpus-based approach?

With regard to the first question, ensuring that the spell checker lexicon has appropriate coverage is crucial to mitigate the over-flagging of correct words as spelling errors and, hence, as non-words for this language, and to improve the correction suggestion lists. Therefore, two of the main tasks have been to define the extent of the required coverage and to facilitate the lexical acquisition process with the instauration of an explicit and consistent procedure that incorporates a number of requirements:

(a) coverage and dialect variation;

(b) neologisms, and borrowings (whether adapted or not); and

(c) the role and status of prescriptive sources.

The issue in (a) is significant when trying to resolve the tension between the considerations of developing one single common tool for the whole Spanish speaking world and the variation inherent to a multinational world language. On one hand, we are building one speller for one language, and, on the other, this speller needs to accommodate all the variation that a multinational language with over 300 million speakers entails. Even though, the spirit of the speller is to be as inclusive as possible, there are factors and logistics limitations that additionally influence the makeup of the coverage. For example, the availability of electronic data plays an important role: data is not homogeneously (quantity and quality-wise) available for all dialects.

The issue in (b) is related to how to deal with the large amount of new words not recorded in dictionaries. Building on the first issue above, it is necessary to take into consideration the fact that Spanish is a language exposed to different language influences (consider the influence of indigenous languages in the Spanish dialects of the Americas, and that of English in many of the Spanish dialects). The issue in (c) relates to the expected behavior of a spell checker in the light of certain recognized prescriptive lexical sources.

Our approach with regard to these factors has been to allow any lexical entry that has been found in recognized and trustful public lexical and textual resources even if prescriptive sources (mainly, the DRAE and the recent DPD) do not include it or include only a spelling variation of it. For instance, the word *closet* has been found in many trustable resources (including the CREA, where the unaccented form outnumbers the accented one) and it has been added to our lexicon along with the prescriptive form *clóset*, both lexical entries are legal forms for our speller. Another example is the case of frequent Latin expressions. Our lexicon includes, for most cases, both the sanctioned spelling (e.g. *a símili*, see DRAE) and the frequent foreign spelling (e.g. *a simili*, cf. DEA).

In relation to the second question on restricted words, the requirement that offensive and inappropriate words be recognized and receive a specific treatment seems to be quite clear: they need to be restricted from appearing as suggestions to misspellings. However, it was not easy to come up with a clear definition of what an offensive or inappropriate word is, not to mention an operational definition. The general criterion that was finally established to deal with this issue was to deem a word offensive only if it was offensive in all of it senses in all the dialects that used that word. Still, there were some cases in which this criterion needed to be bended: extreme cases of words being highly offensive in one dialect, common cases in the targeted main markets, among others.

Regarding the third question, within a corpus-based approach to lexical acquisition and coverage, the role of prescriptive sources is crucial to establish, in conjunction with the notion of frequency, a validation process. Still, there are areas in which the lack of explicitness (and, in lesser degree, the elitism) of the prescriptive norm needed to be overridden by corpus frequency.

[2] It is not the aim of the speller lexicon to cover names extensively. We included only the most common ones. They represent around 1% of the total number of the lexicon entries.

[3] Only a context-based correction technique could detect and correct masking. As already said, masking words are errors that result in another valid word, and it is clear that contextual information is necessary in order to detect and correct them.

5. Evaluation

During the implementation of the speller, around 10K sentences and a list of over 18K misspellings were used to evaluate the speller improvements and regressions. Once the implementation was over, and for evaluation purposes only, a new corpus of 5K sentences was created, of which 3K come from the OWC corpus and 2K from a highly edited corpus. In both cases, the sentences were part of the general corpora reserved for evaluation purposes only.

Data set	3k OWC	2k Edited
# Good Flags	4621	288
# Bad Flags	49	92
# Missed Flags	58	44
# of pages[4]	116	87
# of orthographic errors per page	0.42	1.06
Precision (good/(good + bad))	99%	76%
Recall (good/(good + missed))	99%	87%[5]
False Flags per page (annoyance)	0.42	1.06
Good Flags per page (noticeability)	39.8	3.3

Table 2: Evaluation results.

Table 2 shows how the speller performs on precision (the sum of good and bad flags divided by the good ones) and recall (the sum of good and missed flags divided by the good ones). Following Riley et. al. (2004), false and good flags per page expresses a user experience metric based on the number of good/false flags through the physical concept of pages. False flag per page is perceived from the user's psychological point of view as an annoying experience, while good flags per page are seen as a positive noticeable experience. "# of orthographic errors per page" expresses the result of the number of bad flags divided by the number of pages. These figures put the Spanish speller at the same level of the English one, which is a very remarkable advance.

Table 3 shows suggestion adequacy figures and how the automatic replacement (AR) feature performs. Suggestion adequacy expresses the position of the correction in the suggestion list. Over 91% of the corrections appear in the first five suggestions while the number of correct suggestions appearing in first position is predictably smaller (over 79%), given that the same misspelling can have different corrections based on the intention of the user.

AR performs an automatic replacement of a given misspelling when the first suggestion is considered by far better than the following ones, if any. This feature can be disabled by users. As shown in Table 3, AR triggers 30 times every 100 misspellings and suggestions are 98% of the time correct.

	%
First suggestion	79.77
FirstThroughThird suggestion	89.99
FirstThroughFifth suggestion	91.58
AR Attempt Rate	30.12
AR Success Rate	98.46

Table 3: Suggestion adequacy evaluation results

Although there seems to be a limit of around 80% correction rate for isolated word correction algorithms (Kukich 1992), and a major criticism for this type of technique is that the ranking of alternative correction candidates is fairly imprecise, our evaluation results indicate that categorizing errors and promoting frequent spelling error patterns have had a positive impact in the speller application developed for Spanish.

6. References

[CREA] Real Academia Española: Banco de datos (CREA) [on line]. *Corpus de referencia del español actual*. http://www.rae.es

Damerau, F. (1964) A technique for computer correction of spelling errors. In *Communications of the ACM*, Volume 7, Issue 3, pp. 171-176.

[DEA] Seco, M., Andrés, O. & Ramos, G. (1999). *Diccionario del español actual*. Madrid: Aguilar.

[DPD] Real Academia Española & Asociación de Academias de la Lengua Española (2005). *Diccionario panhispánico de dudas*. Bogotá: Santillana.

[DRAE] Real Academia Española (2003). *Diccionario de la lengua española*. Madrid: Espasa Calpe.

Kukick, K. (1992). Techniques for automatically correcting words in texts. In *ACM Computing Surveys*, 24 (4), pp. 377-439.

Riley, M., Craven, L. & Olsen, M. Customer-focused evaluation of a grammar checker. In Christodoulakis, D. (ed.) (2004): *Pre-Proceedings of the 1st Workshop on International Proofing Tools and Language Technologies*. Patras University, Patras.

Rodríguez, S. & Carretero, J. (1995). Building a Spanish Speller. In *Taller sobre Software de libre distribución*. Universidad Carlos III de Madrid. Spain.

Rodríguez S. & Carretero, J. (1996). A Formal Approach to Spanish Morphology: the COES Tools. In *Sociedad Española para el procesamiento del Lenguaje Natural,* pp. 118-126.

Toole, J. (2000). Categorizing unknown words: using decision trees to identify names and misspellings. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, pp. 173-179.

Veronis, J. (1988). Morphosyntactic correction in natural language interfaces. In *Proceedings of the 12th conference on Computational linguistics,* pp. 708-713.

Vosse, T. (1992). Detecting and correcting morphosyntactic errors in real texts. In *Proceedings of the 3rd Conference Applied Natural Language Processing*, pp. 111-118.

[4] A page is defined according to the following settings: Font: Times New Roman; Font Size: 12pt; Format: one sentence per line (not paragraph format).

[5] This figure is explained because of masking issues, i.e. the text contains contextual misspellings that the speller does not cover.